Locally Developed Oral Skills Evaluation in ESL/EFL Classrooms: A Checklist for Developing Meaningful Assessment Procedures

David N. Ishii and Kyoko Baba

This article explores how teachers, students, and other stakeholders collaboratively develop classroom-based assessment procedures for the evaluation of oral skills. By considering crucial issues in assessment such as validity, teacher-learner collaboration, and contextual factors, the authors provide a checklist that will help ESL/EFL teachers develop meaningful assessment procedures for their own classrooms. The checklist addresses 16 questions worth considering in five test-developing stages: (a) identification of course objectives; (b) identification of skills, strategies, tasks, and content; (c) design of rating procedures; (d) interpretation of learner performance; and (e) reflection on the impact of the assessment procedure. In all the stages the authors emphasize the significance of involving students in the assessment process, which promotes students' responsibility for their own learning.

Cet article analyse la collaboration entre les enseignants, les étudiants et d'autres intervenants dans le développement de procédures d'évaluation des habiletés orales en classe. En tenant compte de questions d'importance cruciale en évaluation telles la validité, la collaboration enseignant-étudiant, et les facteurs contextuels, les auteurs ont dressé une liste de contrôle destinée à aider les enseignants ALS/ALP à développer des procédures d'évaluation qui sont significatives pour leurs propres classes. La liste évoque 16 questions à se poser lors des cinq étapes de développement des évaluations que proposent les auteurs : (a) identification des objectifs du cours ; (b) identification des habiletés, des stratégies, des tâches et du contenu ; (c) conception des procédures d'évaluation ; (d) interprétation de l'apprentissage de l'apprenant et (e) réflexion sur l'impact des procédures d'évaluation. À chaque étape, les auteurs soulignent l'importance d'impliquer les étudiants dans le processus d'évaluation, ce qui favorise la prise en charge par l'étudiant de son propre apprentissage.

Introduction

The move toward more communicative language classrooms has shifted the focus not only of teaching methodologies, but also of assessment approaches. As a result, there is an increasing need for ESL/EFL teachers to refer to a set of criteria when developing, adapting, or adopting oral ability assessments. To serve this need, a growing body of research is being conducted on performance testing, oral proficiency tests in particular, focusing on communicative language use in real situations (Bachman & Palmer, 1996; Clark & Clifford, 1988; Elder, Iwashita, & McNamara, 2002; Hill, 1998; Jenkins & Parra, 2003; Johnson, 2001; Kenyon, 1998; McNamara, 1996; McNamara & Lumley, 1997; Milanovic & Saville, 1996; O'Sullivan, Weir, & Saville, 2002; Patri, 2002; van Lier, 1989). The findings from such research will benefit teachers' assessment practices in language learning classrooms. In this article we are not introducing new concepts. Rather, we have summarized important concepts in the current assessment inquiry so that teachers new to the field may apply them in their own classrooms. By considering relevant issues for the development and use of oral proficiency tests, we present a checklist for ESL/EFL teachers and students to use in collaboratively developing classroom-based assessment approaches for oral skills. We argue that classroom-based assessment can be more meaningful if all the stakeholders involved (i.e., learners, teachers, parents, other community members) negotiate and work together (Burke, 2002; Darling-Hammond, 1994; Darling-Hammond, Ancess, & Falk, 1995).

Collaborative Decision-Making

In the language classroom, who should decide what and how language is assessed? Even though it is commonly expected that the individual instructor (with perhaps institutional guidelines) will make decisions regarding what to teach and how to assess students' language abilities, a question worthy of consideration is should learners be involved in the assessment process? Language instructors know that their classes often take unexpected twists and turns, heading in sometimes unforeseen directions at various moments, often because that learners have their own agendas for learning and assessment in the classroom (Slimani, 1992). Students, especially adult learners, attend classes with their own perspectives on what they would like to learn. As a result, mismatches may occur between the students' needs and the instructional objectives set out by the teacher or institution (Genesee & Upshur, 1996).

For this and other reasons, researchers have espoused the benefits of including learners in the assessment process (Breen, 1989; Breen & Littlejohn, 2000; Ekbatani, 2000; Genesee & Upshur, 1996; Johnson & Johnson, 2002; Shohamy, 2001), a reflection of process-oriented and learner-centered ap-

proaches (Nunan, 1988). What is the justification behind collaborative forms of assessment? To answer this question, we have outlined the reasons for this interest in the form of three metaphors: learners as agents, learners as resources, and learners as evaluators.

Learners as Agents

Learners are active participants who enter classrooms with their own set of attitudes and beliefs about language learning (Cole, 1996; Donato, 2000). Including learners in the decision-making process of assessment increases motivation and fosters positive attitudes because control of the learning process is shared. Learners are, therefore, more apt to be committed or invested in their learning (Johnson & Johnson, 2002) when they engage in collaborative assessment.

Learners as Resources

Learners come with a wealth of background knowledge. They are not empty vessels to be filled with knowledge. Many learners have years of language learning experiences and have developed strategies that may be useful for other learners as well as for the instructor (Cohen, 1998; Johnson & Johnson, 2002). Learners can also provide relevant feedback on how well the assessment procedures are being implemented.

Learners as Evaluators

In a sense, teachers and learners are continually evaluating the successes and failures of a particular task or lesson, even on a daily basis (Breen, 1989). When learners evaluate themselves or others in a meaningful manner, they achieve greater awareness and understanding of both the assessment criteria and their own learning (Ekbatani, 2000; Johnson & Johnson, 2002). If learners act as evaluators in the assessment process, the additional insights gained will strengthen the interpretations of the language performance. We discuss this issue in detail below.

By providing learners with a voice in their assessment process, teachers can improve the quality of learning that takes place in their classrooms. Moreover, Genesee and Upshur (1996) state that "parents, other teachers, noninstructional educational professionals (such as counselors and remedial specialists), and students themselves are also important participants in evaluation" (p. 3). A number of studies have documented the impact of family educational values on academic success (Duran & Weffer, 1992; Kao & Tienda, 1995). Other researchers note that social, institutional, and political support for academic goals is vital to learners' achievement (Cummins, 1996; 2000; Duff, 1995; Taylor, 2001). It is clear that involving the wider community (e.g., parents, counselors, administrators) beyond the classroom may enhance the learning/assessment process.

Understanding the Context

Each language learning context presents its own opportunities and challenges. Before and during instruction, teachers would find it useful to keep in mind several characteristics that influence classroom-based evaluation. Adapting Genesee and Upshur's (1996) "input factors," we have identified and explained the following variables: participants, needs and abilities, motivation and attitudes, prior history, resources, time, and the classroom context.

Participants refers to the characteristics of the learners, such as the number of learners, their first-language backgrounds, proficiency levels, ages, and genders. This information is important because instructors will need to decide whether certain tasks or content are appropriate for their learners. Another aspect of this variable is the identification of other potential participants who may assist in the assessment process. Other teachers, invited experts, or parents may act as raters or as part of the audience, depending on the context and on the types of tasks. For example, two classes may be coordinated such that one class interacts with or gives feedback to the second class, and vice versa.

Often instructors ask students to fill out a quick background questionnaire at the beginning of a course to assist them in obtaining information about their learners' needs and interests. However, the second variable, *needs and abilities*, also reflects the necessity of conducting an ongoing needs analysis and rough inventory of the learners' abilities. Are the needs of the learners changing or becoming more specific? Have their abilities improved to the extent that they are ready to tackle more challenging tasks? The needs and abilities of all other participants, including the teacher, should be identified as well. Is the teacher comfortable with using role-plays or drama in the classroom? Has the invited guest ever conducted mock interviews before?

Motivation and attitudes need to be monitored because these variables change over time. It is impossible for teachers to discern exactly what causes motivation or changes in attitude. Peer-to-peer interactions, factors outside the classroom, technology, or any of a host of other factors may all act as sources of any behavioral or attitudinal changes. However, it may be worthwhile to consider that some conditions are foreseeable. For example, some students may wish to use Powerpoint in their presentations, but its availability may be limited. In this case, will the learners' motivation be sustained if they do not have access to resources that would help them complete the task? In the localized context of a classroom, teachers have the advantage of being able to hear opinions regularly from their students about how they feel about the language learning tasks and motivation.

The *prior histories* of the participants may provide additional help for the teacher. Instructors may decide to ask their students to write brief language

learning histories to elicit information about their previous learning contexts, or they may find it useful to collaborate with the learners' previous instructors to solicit opinions about the students' speaking skills. If learners always expect grammar or teacher-fronted instruction, then the instructor might need to explain the reasons for doing a particular task differently. The instructor may also introduce a new learning technique (such as transcribing) in the classroom that may be unfamiliar to the students. The learners' ability, for example, to transcribe their own tape-recorded speech will depend on their previous experience with transcribing spoken data.

The quality and quantity of *resources* available to the learning context is another variable that affects the learning/assessment process. The design of tasks will inevitably require certain resources that exist in, or are outside, the classroom. Teachers may wish to videorecord their students' role-plays so that they can later reflect on their performances. This may require video cameras, microphones, televisions, adequate space, technical assistance, and other resources. Assigning students the task of interviewing persons for an oral report depends on the availability of such people and their consent.

Teachers and students are continually aware of how much *time* is available to them. Instructors usually want to cover a great deal of content during a course, but must strike a balance between breadth and depth of coverage. Learners often remark that if they only had more time, they would do better. Although instructors may ambitiously decide to ask their students to undertake a number of speaking tasks for assessment purposes, the amount of time required for preparation, practice time, and reflection may not make this feasible.

Finally, the *classroom context* is dynamic in nature. Learners may drop out or new students may enroll, thereby changing the dynamics of the classroom. Target learning goals may become more or less focused. Motivation increases or shows signs of waning for a variety of reasons. Additional resources such as tape recorders may become available to the class. As a result, the context is not a fixed entity. It evolves, just as the teachers and learners do, over time.

Language instructors will benefit from an understanding of those variables specific to their own classroom contexts. These factors are important not only for learning and teaching in general, but also for the evaluation of spoken language. In the following section, we describe a procedure for evaluating oral skills in order to assist teachers in meeting their assessment needs in their specific language learning context.

Model for Oral Skills Evaluation

Taking these factors into account, we propose a five-stage model (see Appendix for checklist) for a locally developed oral skills evaluation (see Figure 1). The components include: (a) identification of course objectives; (b) identifica-

tion of skills, tasks, and content; (c) design of rating procedures; (d) interpretation of learner performance; and (e) reflection on the impact of the assessment procedure. The reciprocal relationship among the components in this model is also reflected in other evaluation frameworks (Bachman & Palmer, 1996; Brown, 1996; Genesee & Upshur, 1996). Evaluation does not proceed in a lock-step manner from one stage to the next. Instead, each stage involves a critical reflection on previous stages as well as that stage's relationship to the changing classroom context.

In each of the five stages, we ask a number of questions for the instructor and learners to consider in order to make the assessment procedure more meaningful and relevant. In order to assist teachers in understanding this procedure, we have stated the reasons for asking these questions. In addition, we have provided sample answers to questions with reference to a specific language learning context: a pre-university, advanced-level, ESL class. Although the ESL classroom used as an example is hypothetical, it is based on our teaching experiences. The Appendix is a summary or checklist of questions for a collaborative approach to oral skills evaluation.

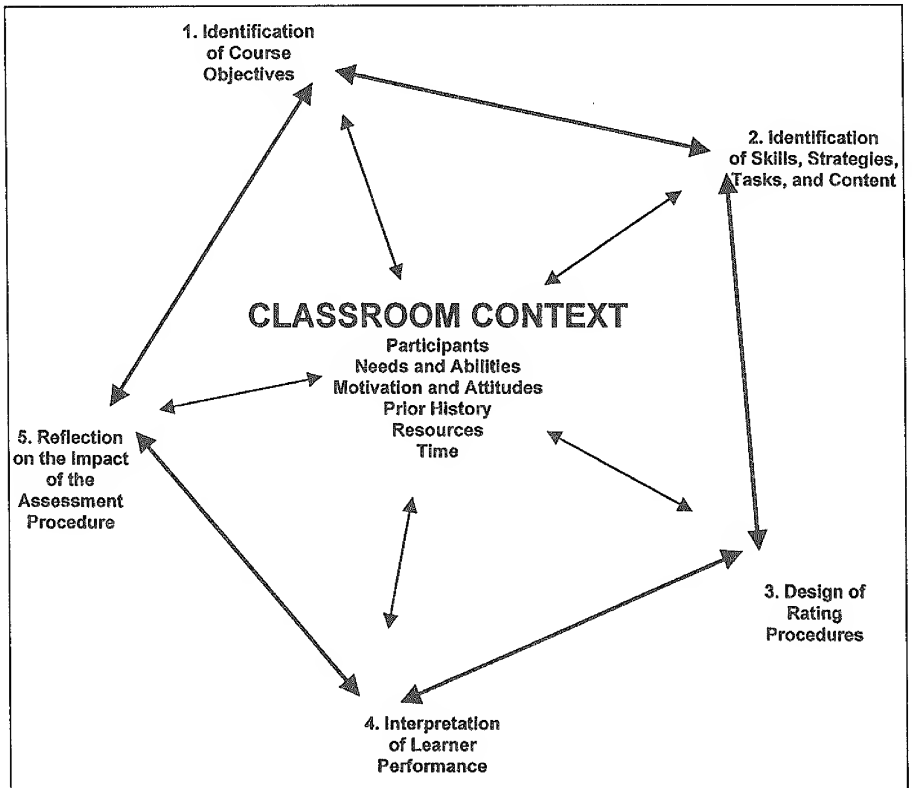

Figure 1. Processes of collaborative oral skills evaluation in classrooms.

Stage 1: Identification of Course Objectives

- In what contexts will the learners speak the target language?
- What kinds of speech are present in the above context?
- Which kinds of speech take priority among the participants?

After referring to the institutional guidelines/course objectives and discussing with or surveying administrators, curriculum designers, and colleagues, instructors may ask themselves and their students to identify the general or specific contexts in which learners will potentially use the language in the immediate and distant future. The identification of contexts will also foster learners' self-reflection on their personal goals for developing oral skills. Where will the learners use the language? In any learning context, students come with a diversity of needs and will therefore respond with a variety of answers.

The instructors and learners next brainstorm the kinds of speech relevant to specific contexts. This helps to identify tasks for both classroom practice and for assessment purposes. Giving presentations, having an interview, talking to customers, holding a conversation are examples of various kinds of speech that might be identified.

With reference to the time constraints of the course, the number of contexts and kinds of speech to focus on in the course will need to be adjusted. Teachers and learners should collaboratively decide which contexts are most relevant to their needs. The instructor can deal with this problem by asking the learners to rank-order the answers listed on the blackboard. The responses may also assist in determining task frequency, sequencing, and difficulty, as well as the weighting of tasks.

In our hypothetical pre-university ESL classroom, instructors openly ask their students where they will use English and what kinds of speech they will need to master. In the academic context, learners may indicate that they need to improve their abilities to give presentations, explain their readings, learn how to ask each other questions in group/class discussions, and/or participate in debates. Authenticity is a crucial characteristic and strength of performance testing (Bachman, 1991; Spolsky, 1985; Wu & Stanfield, 2001). Presentations, for example, are exactly the kinds of speech that students will have to demonstrate in their future studies. However, re-creating the real context is not always possible due to limited resources or time constraints. In that case, it is necessary to specify crucial features of a specific context or characteristic abilities required in that context that can be realized in performance and assessment. At the same time, teachers and students may wish to rank-order which types of speech are most important for their development. Teachers may then introduce a rough syllabus for the course, which could be modified by the student's input (Burke, 2002) in an ongoing manner.

Stage 2: Identification of Skills, Strategies, Tasks, and Content

- What abilities, skills, or strategies are necessary for the students to perform well in the target speech contexts?
- What kind of tasks may be used to assess these skills?
- Are the tasks too difficult or too easy for the learners?
- What topics or content will be used in the tasks?
- Are there any biases?

In order to make the evaluation criteria more explicit, stakeholders need to identify the types of abilities and skills that should be assessed.[1] This will help provide validation evidence for the assessment procedure. Pond, Ul-Haq, and Wade (1995) point out that when evaluation criteria are negotiated between instructors and students, the students clearly understand what should be learned, which in turn fosters increased responsibility for their own learning.

The stakeholders will also need to identify what kinds of tasks (e.g., interviews, presentations, role-plays, group discussion) are appropriate to assess the required skills. The second question asks the teachers, students, and others to consider what kind of interaction will occur in the task.[2] For example, learners will need to understand that the performance of other group members, such as in role-plays, will affect their own oral performance (for more information on using role-plays in assessing oral skills, see Bailey, 1998).

Some learners will naturally find some tasks more difficult than others. Although there are individual differences among learners, it is also important to identify the different task variables (e.g., planning, amount of time-on-task, formality of speech, monologues versus group interaction, etc.) that influence the relative difficulty of the tasks. If more "difficult" speaking tasks are attempted first, then learners may not be able to show their abilities adequately. Task sequencing and grading of tasks must, therefore, be considered during this stage. However, some tasks that were initially perceived as difficult may become easier with instruction and practice time, and some tasks considered difficult by the teacher may not prove to be so for the students.

The selection of appropriate topics for a task is also an important consideration. The purpose of this question is to make the participants realize that topics may be generated by the instructor, the learners themselves, or by participants outside the immediate classroom context. In addition, some topics may be considered irrelevant, insensitive, inappropriate, or taboo. If learners are rated on their oral performance of speaking tasks requiring knowledge of Western culture, then some individuals may have an unfair advantage.[3] Decisions about the choice of topics will depend on what is relevant and interesting to the learners, as well as the course objectives that have been identified earlier.

Not only may some topics favor some individuals, but also some skills and tasks may be biased toward some students. Van Lier (1989) mentions that role-plays, for example, may be biased against shy individuals and may measure acting ability along with speaking skills. This points out the need to gather multiple sources of oral performance data so that learners have a fair chance of displaying their abilities in a variety of tasks (O'Malley & Pierce, 1996; Shohamy, 2001). If the chosen tasks and content are not meaningful and useful to assess the target skills, then perhaps the course objectives in the previous stage will need to be reconsidered.

Referring back to our pre-university academic learning context, all the stakeholders, including learners, may have identified presentations, discussions, and debates as relevant to their learning in Stage 1. Stage 2 involves the identification of specific skills and strategies in these types of speech. For example, the ability to give good presentations requires various skills including clear speech, good eye contact, and appropriate body language. However, the students are likely to experience difficulty in identifying the types of abilities in other forms of speech. Therefore, we recommend that instructors prepare a rough draft of evaluation criteria and review and discuss them with the students. It may also be a good idea to check and modify the draft with other instructors who are teaching (or who have taught) a similar course. Instructors might refer to other resources such as a list of language functions (O'Sullivan, Weir, & Saville, 2002), strategies (Cohen, 1998), or checklists that help teachers/learners provide feedback (Freeman, 1995; Mendelsohn, 1991) to assist them with identifying particular skills.

Stage 3: Design of Rating Procedures

- Who will assess the learners' performance?
- Do the learners know how they will be assessed on the task?
- Under what conditions will the task be performed?

The learners' performance will need to be evaluated by the instructor, by the learners' peers, by the learners themselves, or by individuals outside the classroom. The purpose of the first question is to recognize that the instructor need not be the only person to judge the learners' speaking skills. Genesee and Upshur (1996) state that "actively involving learners in assessing their own progress can sensitize them to instructional objectives and assist them in setting realistic goals for themselves" (p. 45). It should be emphasized that with self- or peer assessment, students learn to manage and regulate their own learning and foster metacognitive skills by reflecting on their performance and critically judging its quality (Ekbatani, 2000; Freeman, 1995; Murphey, 2000; Oscarson, 1989; Patri, 2002; Pond et al., 1995). However, self-/peer-assessment is not without problems, the primary concern being its reliability (Freeman, 1995; MacIntyre, Noels, & Clement, 1997; Patri, 2002). For example, Pond et al. (1995) indicate four problems with peer assessment:

friendship marking (i.e., students give higher marks because they do not want to destroy friendships), collusive marking (i.e., there tends to be no difference between groups), decibel marking (i.e., the noisy performance gets high marks), and parasite marking (i.e., students follow group marks). To avoid these problems, students need to be well trained. The best way to promote students' self-/peer-assessment skills is to make them understand the evaluation criteria clearly and to require them to provide convincing evidence for their marking (Freeman, 1995; Patri, 2002; Pond et al., 1995). Another problem may be that students have negative feelings about peer assessment (e.g., embarrassment, lack of confidence in giving helpful feedback to peers). In order that students benefit from peer assessment, it is important to explain clearly to them the potential advantages of peer assessment (Tang & Tithecott, 1999). This is necessary for the reliability as well as the fairness of the assessment.

Other teachers or invited guests may also provide additional feedback in order to increase reliability of the scores. The question of who will assess the learners' performance also asks the instructor to determine if the rater will play the role of interlocutor, audience, or both. If the rater is an active participant in the speaking task, it is important to recognize that the rater's performance will affect the learner's oral performance. Oral interviews involving only the teacher and learner are examples of tasks in which the learner's speaking skills are dependent on the teacher's performance.

The learners' familiarity with the task instructions, grading criteria, and sample performances is another critical factor for consideration. Raters need to undergo a training process. Although the required skills might have been identified in Stage 2, the grading criteria still need to be fine-tuned. Upshur and Turner (1995) reported how rater teams identified the grading criteria of a speaking task by intuitively sorting out student sample performances into high and low groups. Each group was then subdivided again and again until a set of grading criteria was established. This method can be used collaboratively with the learners so that they may fully understand the assessment process. In the case of writing, instructors have been known to provide sample compositions to give learners an adequate model of performance. However, for oral skills development, the drawback is that the instructor will need to collect a number of sample, tape- or videorecorded performances on a range of tasks for this to be feasible. Instructors may also need to prepare instructions for the raters and learners that explain the rating criteria.

The task conditions of the oral assessment will also affect the learners' oral performances. Where will the task take place? Perhaps the task can be performed outside the classroom. How comfortable is the setting? Is there adequate space, silence, light, air-conditioning? How much time is available to complete the task, and will time constraints affect performance? Will special equipment (e.g., microphones) be used, and will this affect the

learners' performance? Will the raters observe a live performance or an audio/videorecording? Several researchers have suggested that learners could reflect on their speech by examining their own audio/videotapes and transcripts (Allan, 1991; Lynch, 2001; Murphey, 2002). Training will be necessary if such elaborate methods are employed in the classroom.

Once the conditions have been determined, the instructor should provide learners with practice doing the task, practice using the equipment, and practice assessing the performance of themselves and other learners. In addition, the instructor should ensure that all learners are equally assessed under the same conditions. Both the instructors and learners will need to ask themselves if the rating procedures have been established so that fair ratings will be obtained.

In our sample ESL academic context, the instructor may want to ask a colleague to assess students' performances independently in order to increase the reliability of instructors' ratings. Moreover, the teacher and learners will need to discuss the option of including self- or peer assessments. Multiple raters will assist in strengthening the interpretations of oral performances. The teachers and students could agree that the weighting of the ratings, for example, could be: instructor's evaluation 60%, peer assessment 30%, and self-assessment 10%. In the case of academic presentations, learners could assess each other using the evaluation criteria identified in Stage 2. In addition, teachers may wish to encourage learners to reflect on their own progress and to assess their own presentations by watching videotapes. If no video equipment is available, students could be asked to evaluate their own presentations through immediate recall.

Stage 4: Interpretation of Learner Performance

- Are the raters consistent with their ratings and with other raters?
- How are the results simplified for administrative purposes?
- What kind of and how much feedback should be provided to the learners?
- Is the performance assessed with the target abilities in mind?

After the learners perform a task, the instructors and learners will need to interpret the scores or ratings. One issue that might arise during the interpretation is a discrepancy in the ratings. If the learners' performance is rated by a number of individuals, the ratings may show a wide variation. The instructor at this point may need to negotiate whether to discuss the ratings and if a consensus needs to be reached. For example, some raters may overemphasize certain skills (e.g., grammatical accuracy), which would compromise the inferences made about the oral abilities of the learner.

Another issue relates to the raters' consistency with their own ratings. Raters may become more harsh or lenient in their grading over time due to tiredness or other factors. This may be alleviated by ensuring that the raters

have sufficient breaks in between ratings. In addition, raters should have time to go over their marks at the end to ensure that there is no unfair fluctuation in scoring.

At the end of the course, the instructor will inevitably report the results for administrative purposes. How will this be accomplished? Although the institution may already have decided the accepted form in which scores are reported (e.g., letter grades, percentages, pass/fail), the instructor will need to consider how to simplify the learners' performances of specific tasks into a reportable form.

The reporting of results should benefit the learners as well. Instructors should consider how to provide meaningful feedback to promote learner development and to show progress over time (e.g., Hyland, 2003; Sze, 2002). Instructors and learners can employ a combination of methods including holistic/discrete point scoring, detailed descriptions of learner abilities, error corrections, and/or suggestions for further improvement. People often form varied impressions of the same type of feedback (Kouritzin & Vizard, 1999). We recommend that students receive different types of feedback on their performances (e.g., videotaping students, audiotaped comments from the teacher, written feedback).

The instructor and learners should also confirm that the assessment targets the abilities (e.g., linguistic and nonlinguistic abilities, etc.) that were previously agreed on in Stage 2. If the ratings reflect skills and abilities other than those identified, then perhaps the grading criteria need to be revised. The reporting of the results should reflect what is appropriate and meaningful for learners and other users of the test scores. The instructors and learners can ask themselves whether the assessment enables them to make adequate and appropriate inferences about the learner' performance.

In a pre-university ESL context, the ordering of students' oral presentations is an example where inconsistencies may occur, because the initial presenter may be graded too severely or too lightly. In such a case, raters should be given time to review their marks and comments at the end of the presentations. How can the reliability of the assessment be achieved in other types of tasks? Even if the students have undertaken practice in self- or peer assessment, the results are still subjective. Because most students participate in this ESL program voluntarily and are not obliged to submit the grading to institutions outside the classroom context, the assessment is formative and directed toward helping language learners perform better on high-stakes gatekeeping exams. That is, the assessment mainly provides students with immediate feedback to promote their learning, a goal instructors can share with their students. Periodic one-on-one conferences or interviews with the students may be especially beneficial to the students because they can carefully examine all the comments they have received, summarize progress toward their goals, and clarify their aims with the instructor.

Stage 5: Reflection on the Impact of the Assessment Procedures

- Were the procedures and results of the assessment meaningful for both the instructors and learners?
- Will the administration of the assessment itself change teaching and learning in a positive way?

Teachers and learners will continually evaluate and reflect on their teaching and learning throughout the assessment process. However, there is a need for a final reflection on the impact of the assessment process (with regard to *washback*, see, e.g., Bailey, 1999; Wall, 2000; Wall & Alderson, 1993). Some learners, for example, will have achieved a greater sense of awareness of the grading criteria for successful oral performance. Others may have noticed the amount of progress that they have made over the course of the term. It is also conceivable that the assessment process has an impact on instructors, curriculum designers, or administrators. Based on the results of the assessment process, instructors may change their methods of teaching and assessment. Curriculum designers may modify or improve the present curriculum by examining the feedback from instructors. Administrators may reconsider the quantity and quality of the students they accept to their institution, or change the class size and placement decision procedures.

Instructors may also realize the need to increase the amount of in-class practice time of speaking activities. Bygate (1987) makes reference to exploratory and final draft learning. Final draft learning refers to the learners' aim of producing target-like, error-free speech, whereas exploratory learning indicates the learners' need to experiment with language in spoken discourse. The assessment procedures should provide room for both types of learning. If instructors employ an assessment procedure without much practice or exploratory time, they are denying learners the opportunity to take risks by trying out new vocabulary, pronunciation, or ways of speaking.

In our hypothetical ESL context, the instructor can try to elicit student opinions regarding learning and assessment through questionnaires, group/class discussions, or individual conferences. Although the impact of the assessment procedures may normally be reviewed at the end of the course, instructors may perhaps consider addressing this issue midway through the course as well. The summary of the comments from the students and the instructor's reflections could be reported to other instructors and administrators, who may also adjust teaching and assessment guidelines.

Concluding Remarks

In summary, we hope this framework promotes a systematic method for collecting multiple sources of evidence of oral language ability in context-specific learning environments. The use of a variety of assessment procedures will assist in providing more valid measures of oral ability. The

collaborative nature of the evaluation process will also allow learners and other stakeholders to have a voice in the learning process, thereby increasing the investment of all participants. This could have a positive effect on the teaching and learning of oral skills in the second-language classroom.

We outline numerous questions for the instructor to consider. We understand that collaboration at all levels of the assessment process requires much time and effort. However, it is our hope that instructors will include their learners in some aspects of the assessment process. The close relationship between the teacher and students is a feature of classroom-based assessment. It is in the classroom that they can collaborate and locally develop their own assessment procedures to suit their own specific needs for language learning. This may change the instructors' and learners' perceptions so that they may regard assessment as a learning experience instead of a gatekeeping or judgment experience.

Notes

[1]Although many instructors may understand that both listening comprehension and speech production are components of oral skills, it is necessary to recognize that successful communicative performance in oral tasks also hinges on a broader set of interrelated abilities. This includes linguistic skills (e.g., pronunciation, grammar, vocabulary), nonlinguistic abilities (e.g., eye contact, body language), affective factors (e.g., attitude, motivation), interaction skills (e.g., the ability to negotiate), and cognitive skills (e.g., memory, planning, McNamara, 1996).

[2]Successful communication not only relies on a set of interrelated abilities, but it also depends on the behavior and abilities of the other individual during a conversational interaction. Even in occasional monologic speech such as in presentations, the attentiveness of the audience will affect the presenter's performance. A rating of a conversation between two individuals (e.g., teacher-student, learner-learner) is thus a rating of the interaction and not of one particular person (McNamara, 1997). Learners may find it naturally easier to talk to some individuals than others (Storch, 2002). Although people may attribute this to personality differences, the concept of oral performance as an interaction needs to be considered in the assessment of oral skills (Hall, 1995; Kramsch, 1986; Young, 1999). From an evaluative point of view, this further supports the need to obtain multiple sources of evidence and to collect them from a variety of sources including teacher, peer, and self-assessments (Chalhoub-Deville, 1995).

[3]Fairness is a comprehensive concept that should be considered throughout the entire process of test development and use. Refer to Kunnan (2000) and International Language Testing Association (2000) for further information.

Acknowledgments

We thank Alister Cumming, Lindsay Brooks, and *TESL Canada Journal*'s anonymous reviewers for their insightful comments on earlier drafts of our article.

Correspondence concerning this article may be addressed to David Ishii, Modern Language Center, Ontario Institute for Studies in Education, University of Toronto, 252 Bloor Street West, Toronto, Ontario, Canada M5S 1V6. E-mail address: dishii@oise.utoronto.ca.

The Authors

David N. Ishii and Kyoko Baba are doctoral candidates in the Ontario Institute for Studies in Education at the University of Toronto.

References

Allan, D. (1991). Tape journals: Bridging the gap between communication and correction. *Journal, 45*, 61-66.

Bachman, L.F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*, 671-704.

Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Bailey, K.M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*, 257-279.

Bailey, K. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. New York: International Thomson.

Bailey, K. (1999). *Washback in language testing*. TOEFL Monograph Series. Princeton, NJ: Educational Testing Service.

Breen, M. (1989). The evaluation cycle for language learning tasks. In R.K. Johnson (Ed.), *The second language curriculum* (pp. 187-206). New York: Cambridge University Press.

Breen, M., & Littlejohn, A. (2000). The significance of negotiation. In M. Breen & A. Littlejohn (Eds.), *Classroom decision-making: Negotiation and process syllabuses in practice* (pp. 5-39). New York: Cambridge University Press.

Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Burke, R. (2002). Formative assessment procedures and the second language curriculum: Signposts for the journey. *TESL Canada Journal, 19*(2), 87-91.

Bygate, M. (1987). *Speaking*. New York: Oxford University Press.

Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning, 45*, 251-281.

Clark, J., & Clifford, R. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques. *Studies in Second Language Acquisition, 10*, 129-147.

Cohen, A.D. (1998). *Strategies in learning and using a second language*. London: Longman.

Cohen, A.D. (2000). Exploring strategies in test-taking: Fine-tuning verbal reports from respondents. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp. 127-150). Mahwah, NJ: Erlbaum.

Cole, M. (1996). *Cultural psychology: A once and future discipline*. Cambridge, MA: Belknap Press.

Cummins, J. (1996). *Negotiating identities: Education for empowerment in a diverse society*. Los Angeles, CA: California Association for Bilingual Education.

Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon, UK: Multilingual Matters.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64*(1), 5-30.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. New York: Teachers College Press.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, UK: Cambridge University Press.

Donato, R. (2000). Sociocultural contributions to understanding the foreign and second language classroom. In J. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 27-50). Oxford, UK: Oxford University Press.

Duff, P. (1995). An ethnography of communication in immersion classrooms in Hungary. In K. Davis & A. Lazarton (Eds.), Qualitative research in ESOL [Special issue]. *TESOL Quarterly, 29*, 505-538.

Duran, B., & Weffer, R. (1992). Immigrants' aspirations, high school process, and academic outcomes. *American Educational Research Journal, 1*, 163-181.

Ekbatani, G. (2000). Moving toward learner-directed assessment. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp. 1-11). Mahwah, NJ: Erlbaum.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 18*, 347-368.
Freeman, M. (1995). Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education, 20*(3), 289-300.
Genessee, F., & Upshur, J. (1996). *Classroom-based evaluation in second language education*. New York: Cambridge University Press.
Hall, J. (1995). (Re)creating our worlds with words: A sociocultural perspective of face-to-face interaction. *Applied Linguistics, 16*, 145-166.
Hill, K. (1998). The effect of test-taker characteristics on reactions to and performance on an oral English proficiency test. In A. J. Kunnan (Ed.), *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 209-229). Mahwah, NJ: Erlbaum.
Hyland, F. (2003). Focusing on form: Student engagement with teacher feedback. *System, 31*, 217-230.
International Language Testing Association. (2000). *Code of ethics for ILTA*. Adopted at the annual meeting of the International Language Testing Association, Vancouver.
Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *Modern Language Journal, 87*, 90-107.
Johnson, D., & Johnson, R. (2002). *Meaningful assessment: A manageable and cooperative process*. Boston, MA: Allyn and Bacon.
Johnson, M. (2001). *The art of non-conversation: A reexamination of the validity of the oral proficiency interview*. New Haven, CT: Yale University.
Kao, G., & Tienda, M. (1995). Optimism and achievement: The educational performance of immigrant youth. *Social Science Quarterly, 76*(1), 1-19.
Kenyon, D.M. (1998). An investigation of the validity of task demands on performance-based tests of oral proficiency. In A. J. Kunnan (Ed.), *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 19-40). Mahwah, NJ: Erlbaum.
Kouritzin, S.G., & Vizard, C. (1999). Feedback on feedback: Preservice ESL teachers respond to evaluation practices. *TESL Canada Journal, 17*(1), 16-39.
Kramsch, C. (1986). From language proficiency to interactional competence. *Modern Language Journal, 70*, 366-371.
Kunnan, A.J. (1998). Approaches to validation in language assessment. In A.J. Kunnan (Ed.), *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 1-16). Mahwah, NJ: Erlbaum.
Kunnan, A. (2000). Fairness and justice for all. In A. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th language testing research colloquium, Orlando, Florida* (pp. 1-14). Cambridge: Cambridge University Press.
Lynch, T. (2001). Seeing what they meant: Transcribing as a route to noticing. *ELT Journal, 55*, 124-132.
MacIntyre, P.D., Noels, K.A., & Clement, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning, 47*, 265-287.
McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.
McNamara, T. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics, 18*, 446-466.
McNamara, T.F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*, 140-156.
Mendelsohn, D. (1991). Instruments for feedback in oral communication. *TESOL Journal*, 25-30.

Milanovic, M., & Saville, N. (1996). Introduction. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th language testing research colloquium, Cambridge and Arnhem* (pp. 1-17). Cambridge, UK: Cambridge University Press.

Murphey, T. (2002). Videoing conversations for self-evaluation in Japan. In J. Murphy & P. Byrd (Eds.), *Understanding the courses we teach: Local perspectives on English language teaching* (pp. 179-196). Ann Arbor, MI: University of Michigan Press.

Nunan, D. (1988). *The learner-centred curriculum.* New York: Cambridge University Press.

O'Malley, J., & Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers.* Reading, MA: Addison Wesley.

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6,* 1-13.

O'Sullivan, B., Weir, C.J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19,* 33-56.

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing, 19,* 109-131.

Pond, K., Ul-Haq, R., & Wade, W. (1995). Peer review: A precursor to peer assessment. *Innovation in Education and Training International, 32,* 314-323.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests.* New York: Addison Wesley Longman.

Slimani, A. (1992). Evaluation of classroom interaction. In J.C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 197-221). New York: Cambridge University Press.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2,* 31-40.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52,* 119-158.

Sze, C. (2002). A case study of the revision process of a reluctant ESL student writer. *TESL Canada Journal, 19*(2), 21-36.

Tang, G.M., & Tithecott, J. (1999). Peer response in ESL writing. *TESL Canada Journal, 16*(2), 20-38.

Taylor, S. (2001). *Triangualism by design? An investigation into the educational experience of Kurdish children schooled in Denmark.* Unpublished doctoral dissertation, Ontario Institute for Studies in Education of the University of Toronto.

Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal, 49,* 3-12.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly, 23,* 489-508.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System, 28,* 499-509.

Wall, D., & Alderson, J.C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10,* 41-69.

Wu, W.M., & Stanfield, C.W. (2001). Towards authenticity of task in test development. *Language Testing, 18,* 187-206.

Young, R. (1999). *Sociolinguistic approaches to SLA: Annual Review of Applied Linguistics, 19,* 105-131.

Appendix
Checklist for Locally Developed Oral Skills Evaluation

1. *Identification of Course Objectives*

- In what contexts will the learners speak the target language? (e.g., education, employment, home, etc.)
- What kinds of speech are present in the above context? (e.g., giving presentations, talking to customers, holding a conversation, etc.)
- Which kinds of speech take priority among the participants?

2. *Identification of Skills, Strategies, Tasks, and Content*

- What abilities, skills, or strategies are necessary to perform well in the target kinds of speech? (e.g., giving presentations—clear speech, eye contact, body language, etc.)
- What kind of tasks may be used to assess these skills? (e.g., interview, presentation, group discussion, debates, etc)
- Are the tasks too difficult or too easy for the learners?
- What topics or content will be used in the task? (in addition, are the topics teacher-, student-, or other-generated?)
- Are there any potential biases? (e.g., cultural or gender biases, background knowledge, personality, etc.)

3. *Design of Rating Procedures*

- Who will assess the learners' performance? (e.g., instructors, peer- and self- assessments, invited guests)
- Do the learners know how they will be assessed on the task? (i.e., are the raters familiar with the task instructions, grading criteria, sample performances?)
- In what conditions will the task be performed? (e.g., setting, time constraints, special equipment, etc.)

4. *Interpretation of Learner Performance*

- Are the raters consistent with their ratings and with other raters?
- Is the performance assessed with the target abilities in mind?
- How are the results simplified for administrative purposes? (e.g., letter grade, percentages, pass/fail)
- What types and how much feedback should be provided to the learners? (e.g., holistic/ discrete point scoring, detailed description of learner abilities error corrections, suggestions for further improvement)

5. *Reflection on the Impact of the Assessment Procedures*

- Will the administration of the assessment itself change your teaching and learning in a positive way? (e.g., increased in-class practice time on speaking activities, greater awareness of the grading criteria for successful oral performance, etc.)